

ECE 544

11/13/2007

How to choose a classifier

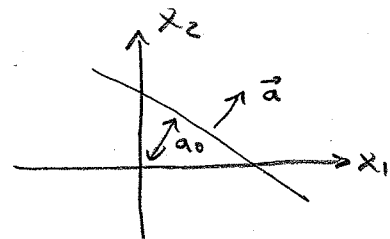
TODAY

1. VC Dimension & PAC Bounds
2. No Free Lunch
3. Bias & Consistency
4. Bayesian Model Selection & MDL

VC Dimension

EXAMPLE: 2D LINEAR CLASSIFIER

$$\alpha(\vec{x}) = \text{sign}(\vec{a}^T \vec{x} + a_0)$$



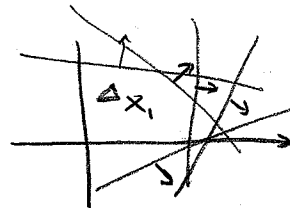
How many different class boundaries can it represent?

Definition of "Different"

$n=1: \mathcal{D} = \{\vec{x}_1\} \Rightarrow 2$ different boundaries

$$\vec{a}^T \vec{x}_1 + a_0 < 0$$

$$\vec{a}^T \vec{x}_1 + a_0 > 0$$

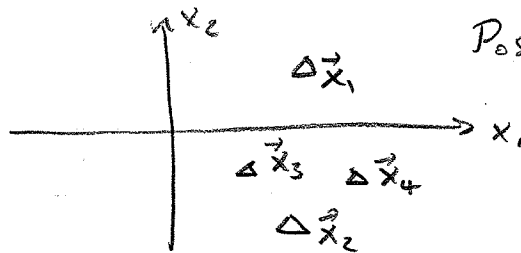


Equivalent classifiers
 $\vec{a}^T \vec{x}_1 + a_0 < 0$

$n=2: \mathcal{D} = \{\vec{x}_1, \vec{x}_2\} \Rightarrow 4$ different classifiers

$n=3: 8$ different classifiers

$n=4: ?$



Possible decision regions \mathcal{R}_{+1} :

- $\mathcal{R}_1 = \emptyset, \{\vec{x}_2\}, \{\vec{x}_1\}, \{\vec{x}_3\}, \{\vec{x}_4\},$
- $\{\vec{x}_2, \vec{x}_3\}, \{\vec{x}_2, \vec{x}_4\}$
- but not $\{\vec{x}_3, \vec{x}_4\}!!$

$N_p(n) = \#$ possible partitions of n -sample training set

Linear 2-D classifier:

- ① Rotating \vec{x} gives n data orderings
 - ② Threshold each ordering $\Rightarrow n$ partitions/orderings
- $\Rightarrow N_p(n) \leq n^2$

Linear d-Dimensional classifier: $N_p(n) \leq n^d$

Vapnik-Chervonenkis Dimension of a classifier:

$$d_{vc} \equiv \lim_{n \rightarrow \infty} \frac{\log N_p(n)}{\log n}$$

Probably Approximately Correct Bounds (PAC Bounds)

Define $L_\alpha(x_i, y_i) = \text{loss function, e.g., } L = [\alpha(x_i) \neq y_i]$
 $R(\alpha) = E[L_\alpha] = \iint L_\alpha(x, y) p(x, y) dx dy = \text{Risk}$

$$R_{\text{emp}}(\alpha, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n L_\alpha(x_i, y_i) = \text{"Empirical Risk"}$$

Assume (x_i, y_i) i.i.d. selections from $p(x, y)$

Then

$$P\left(|R(\alpha) - R_{\text{emp}}(\alpha)| < G\left(\frac{d_{vc}}{n}, \delta\right)\right) > \underbrace{1 - \delta}_{\text{"confidence"}}$$

$$G\left(\frac{d_{vc}}{n}, \delta\right) = \text{"Generalization Error"}$$

$$G\left(\frac{d_{vc}}{n}, \delta\right) \approx \frac{d_{vc}}{n} \quad (\text{only weak dependence on } \delta)$$

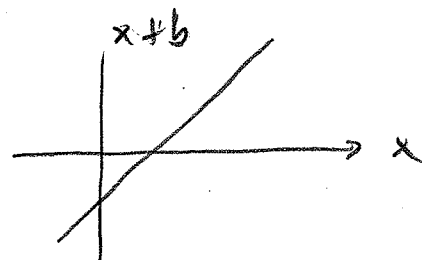
$$\Rightarrow \boxed{R(\alpha) < R_{\text{emp}}(\alpha) + G\left(\frac{d_{vc}}{n}\right) \text{ w/ prob } 1 - \delta}$$

Instructive Examples

$$x = \text{scalar}, \quad y \in \{-1, 1\}$$

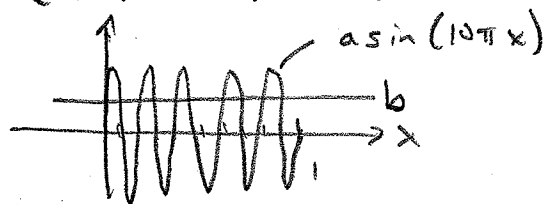
$$\text{Classifier 1: } \alpha(x) = \text{sign}(x+b)$$

$$d_{vc} = 1.$$



$$\text{Classifier 2: } \alpha(x) = \text{sign}(\sin(10\pi x) + b)$$

$$d_{vc} = 1$$



Which is better?

A: Whichever produces lower $R_{emp}(\alpha)$

\approx whichever better matches reality!

No Free Lunch Define $\text{Ave}(f) = \text{average over all } p(x, y)$

Version 1 (Textbook): \mathcal{D} not i.i.d.

For any particular $p(x, y)$,
 \exists as many "good" \mathcal{D} as "bad" \mathcal{D} ,
thus $\forall \mathcal{D}$ selected from $p(x, y)$ without i.i.d. assumption,

$$\text{Ave}(R(\alpha_1)) - \text{Ave}(R(\alpha_2)) = 0$$

regardless of $R_{emp}(\alpha_1), R_{emp}(\alpha_2)$
for any classifiers α_1, α_2

Version 2 (Not in textbook): \mathcal{D} i.i.d.

$$\text{Given } R_{emp}(\alpha_1) = R_{emp}(\alpha_2)$$

$$P(R(\alpha_1) < R(\alpha_2)) = \frac{1}{2} \quad \text{independent of } d_{vc}(\alpha_1), d_{vc}(\alpha_2)$$

Reason: PAC bounds tell us that, w/prob $1-\delta$

$$R_{\text{emp}} - \frac{\text{dvc}(\alpha_1)}{n} < R(\alpha_1) < R_{\text{emp}} + \frac{\text{dvc}(\alpha_1)}{n}$$

$$R_{\text{emp}} - \frac{\text{dvc}(\alpha_2)}{n} < R(\alpha_2) < R_{\text{emp}} + \frac{\text{dvc}(\alpha_2)}{n}$$

$\text{dvc}(\alpha_1) < \text{dvc}(\alpha_2) \Rightarrow R(\alpha_1)$ has low variance
but mean is same

BIAS & CONSISTENCY

Let $g(\vec{x}_0 | \mathcal{D}) =$ function of \vec{x}_0 , trained on $\mathcal{D} = \{\vec{x}_1, \dots, \vec{x}_n\}$

$$\text{Examples: } \textcircled{1} \quad g(\vec{x}_0 | \mathcal{D}) = \frac{1}{(2\pi)^{d/2} |\Sigma(\mathcal{D})|^{1/2}} e^{-\frac{1}{2}(\vec{x}_0 - \mu(\mathcal{D}))^T \Sigma(\mathcal{D})^{-1} (\vec{x}_0 - \mu(\mathcal{D}))}$$

Gaussian
trained on \mathcal{D}

$$\mu(\mathcal{D}) \equiv \frac{1}{n} \sum_{i=1}^n \vec{x}_i$$

$$\Sigma(\mathcal{D}) \equiv \frac{1}{n-1} \sum_{i=1}^n (\vec{x}_i - \mu(\mathcal{D})) (\vec{x}_i - \mu(\mathcal{D}))^T$$

$$\textcircled{2} \quad g(\vec{x}_0 | \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(\vec{x}_0 - \vec{x}_i)$$

— Parzen window
estimate

$$\textcircled{3} \quad g(\vec{x}_0 | \mathcal{D}) = \text{neural net trained on } \mathcal{D} \\ \text{starting from known } \vec{w}_0$$

$$\text{MSE}(\vec{x}_0) = E \left[(g(\vec{x}_0 | \mathcal{D}) - F(\vec{x}_0))^2 \right]$$

↑ Target function

$$= \iiint (g(\vec{x}_0 | \mathcal{D}) - F(\vec{x}_0))^2 p(\vec{x}_1) \dots p(\vec{x}_n) d\vec{x}_1 \dots d\vec{x}_n$$

Then

$$\text{MSE}(\vec{x}_0) = \text{BIAS}^2 - \text{VAR}(g)$$

$$\text{BIAS} \equiv E_{\mathcal{D}} [g(\vec{x}_0 | \mathcal{D})] - F(\vec{x}_0)$$

$$\text{VAR}(g) \equiv E_{\mathcal{D}} \left[(g(\vec{x}_0 | \mathcal{D}) - E[g(\vec{x}_0 | \mathcal{D})])^2 \right]$$

Low bias $\Leftrightarrow g(\vec{x}_0 | \mathcal{D})$ flexible (high dvc)

Low variance $\Leftrightarrow g(\vec{x}_0 | \mathcal{D})$ not too dependent on \mathcal{D}
(low dvc)

EXAMPLE: CLASSIFICATION

Suppose $g(\vec{x}_0 | \mathcal{D}) = \alpha(\vec{x}_0) \in \begin{cases} w_1 & \text{class 1} \\ w_0 & \text{class } \emptyset \end{cases}$

Define $F(\vec{x}_0) \equiv P(y_0 = w_1 | \vec{x}_0)$

Zero-Bias Classifier

Suppose $\alpha(\vec{x}_0)$ good enough so $E[g(\vec{x}_0 | \mathcal{D})] = F(\vec{x}_0)$

$$\begin{aligned} \text{Then } \text{VAR}(g) &= P(w_1 | \vec{x}_0) (1 - P(w_1 | \vec{x}_0))^2 \\ &\quad + (1 - P(w_1 | \vec{x}_0)) (0 - P(w_1 | \vec{x}_0))^2 \end{aligned}$$

$$= P(w_1 | \vec{x}_0) (1 - P(w_1 | \vec{x}_0))$$

Zero-Variance Case

Suppose $g(\vec{x}_0 | \mathcal{D}) = 1$ everywhere!

$$\begin{aligned} \text{Then } \text{BIAS}^2 &= (1 - F(\vec{x}_0))^2 = (1 - P(w_1 | \vec{x}_0))^2 \\ &= \text{MSE}(\vec{x}_0) \end{aligned}$$

4. BAYESIAN CLASSIFIER SELECTION & MINIMUM DESCRIPTION LENGTH

GOAL: Choose $\alpha(\vec{x}_0 | \mathcal{D})$ to minimize

$$\log P(\alpha, \mathcal{D}) = \log P(\alpha) + \log P(\mathcal{D} | \alpha)$$

$\log P(\mathcal{D} | \alpha) = \log$ likelihood of data

$\log P(\alpha) =$ prior over "reasonable" hypotheses

Suppose $\alpha \in A$, the set of hypotheses

Consider a Huffman code $C(\alpha) = [c_1, c_2, \dots]$

$c_i =$ i TH BIT

$$c_1 = 0 \quad \text{IF } \alpha \in A_0 \quad P(\alpha \in A_0) = \frac{1}{2}$$

$$c_1 = 1 \quad \text{IF } \alpha \in A_1 \quad P(\alpha \in A_1) = \frac{1}{2}$$

$$[c_1, c_2] = [0, 1] \quad \text{IF } \alpha \in A_{01}, \quad P(\alpha \in A_{01}) = \frac{1}{4}$$

\vdots

Then $\text{length}(C(\alpha)) = \log_2 P(\alpha)$ bits

Similarly let $C(\mathcal{D} | \alpha) = [d_1, d_2, \dots]$

$$[d_1, d_2] = [0, 1] \quad \text{IF } \mathcal{D} \in \mathcal{R}_{01}, \quad P(\mathcal{D} \in \mathcal{R}_{01} | \alpha) = \frac{1}{4}$$

Then $\text{length}[C(\mathcal{D} | \alpha)] = \log_2 P(\mathcal{D} | \alpha)$

\Rightarrow Choose α to minimize description length

$$\text{length}[C(\alpha)] + \text{length}[C(\mathcal{D} | \alpha)]$$

Example, d -Dimensional hyperplane: $\text{length}[C(\alpha)] \propto d+1$